EURASIP Journal on Applied Signal Processing 2003:3, 264–275

Improved Facial-Feature Detection for AVSP via Unsupervised Clustering and Discriminant Analysis

Simon Lucey
Speech Research Laboratory, RCSAVT, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia
Email: slucey@ieee.org

Sridha Sridharan
Speech Research Laboratory, RCSAVT, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia
Email: s.sridharan@qut.edu.au

Vinod Chandran
Speech Research Laboratory, RCSAVT, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia
Email: v.chandran@qut.edu.au

Received 22 February 2001 and in revised form 21 June 2002

An integral part of any audio-visual speech processing (AVSP) system is the front-end visual system that detects facial features (e.g., eyes and mouth) pertinent to the task of visual speech processing. The ability of this front-end system to not only locate, but also give a confidence measure that the facial feature is present in the image, directly affects the ability of any subsequent postprocessing task such as speech or speaker recognition. With these issues in mind, this paper presents a framework for a facial-feature detection system suitable for use in an AVSP system, but whose basic framework is useful for any application requiring frontal facial-feature detection. A novel approach for facial-feature detection is presented, based on an appearance paradigm. This approach, based on intraclass unsupervised clustering and discriminant analysis, displays improved detection performance over conventional techniques.

Keywords and phrases: audio-visual speech processing, facial-feature detection, unsupervised clustering, discriminant analysis.

1. INTRODUCTION

The visual speech modality plays an important role in the perception and production of speech. Although not purely confined to the mouth, it is generally agreed [1] that the large proportion of speech information conveyed in the visual modality stems from the mouth region of interest (ROI). To this end, it is imperative that an audio-visual speech processing system be able to accurately detect, track, and normalise the mouth of a subject within a video sequence. This task is referred to as *facial-feature detection* (FFD) [2]. The goal of FFD is to detect the presence and location of features, such as eyes, nose, nostrils, eyebrows, mouth, lips, ears, and so on, with the assumption that there is *only* one face in an image. This differs slightly from the task of facial-feature location which assumes that the feature is present and only requires its location. Facial-feature tracking is an extension to the task of location in that it incorporates temporal information in a video sequence to follow the location of a facial feature as time progresses.

The task of FFD, with reference to an audio-visual speech processing (AVSP) application, can be broken into three parts, namely,

(1) the initial location of a facial-feature search area at the beginning of the video sequence;
(2) the initial detection of the eyes at the beginning of the video sequence. Detection is required here to ensure that the scale of the face is known for normalisation of the mouth in the AVSP application;
(3) the location and subsequent tracking of the mouth throughout the video sequence.

A depiction of how the FFD system acts as a front-end

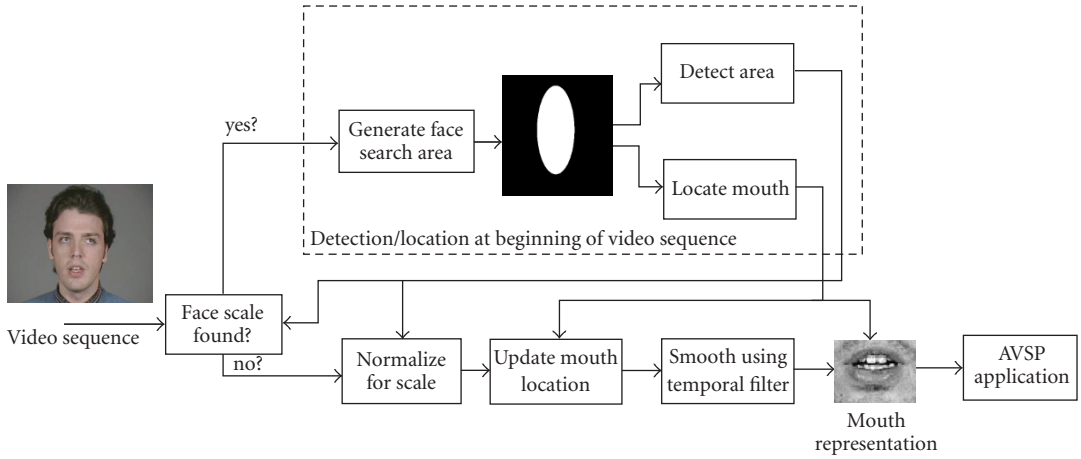


Figure 1: Graphical depiction of overall detection/location/tracking front-end to an AVSP application.

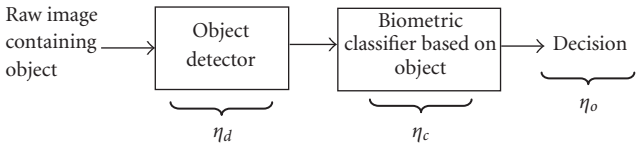


Figure 2: Graphical depiction of the cascading front-end effect.

to an AVSP application can be seen in Figure 1. This paper is broken down into a number of sections. Firstly, Section 2 discusses the importance of the front-end FFD system has on the overall performance of an AVSP application. Section 3 discusses the scope of the FFD problem with reference to AVSP and how some assumptions can be made to simplify the system (i.e., lighting, number of people present, scale and rotation of face, and so on). Under these assumptions, a technique for generating a binary face map, to restrict the eye and mouth search space, is explained in Section 5. The importance of the face map can be seen in Figure 1 as it can drastically reduce the search space in FFD. In Section 6, an appearance-based paradigm for FFD is defined, with our new approach of detection based on intraclass unsupervised clustering and discriminant analysis being outlined. Detection results of this approach highlighting the improved performance attained over conventional techniques are also presented.

2. FRONT-END EFFECT

For biometric processing of the face, it is common practice to perform *manual* labelling of important facial features (i.e., mouth, eyes, etc.) so as to remove any bias from the *front-end effect*. The *front-end effect* can be defined as the dependence any visual biometric classifier's performance has on having the object it is making a decision about accurately detected. The severe nature of this effect, with reference to final biometric performance, is best depicted in Figure 2.

If we assume that an erroneous decision will result when the facial feature being classified is not successfully detected, we can mathematically express the effect as

$$\eta_o = \eta_d \times \eta_c, \tag{1}$$

where η_d is the probability that the object has been successfully detected, η_c is the probability that a correct decision is made, given that the object has been successfully detected, and η_o is the overall probability that the system will make the correct decision. Inspecting (1), we can see that the performance of the overall classification process η_o can be severely affected by the performance η_d of the detector.

In ideal circumstances, we want η_d to approach unity, so we can concentrate on improving the performance of η_c, thus improving the overall system performance. A very simple way to ensure η_d approaches unity is through manual labelling of facial features. Unfortunately, due to the amount of visual data needing to be dealt with in an AVSP application, manual labelling is not a valid option. The requirement for manually labelling facial features also brings the purpose of any automatic classification system (i.e., speech or speaker recognition) into question due to the need for human supervision. With these thoughts in mind, an integral part of any AVSP application is the ability to make η_d approach unity via an automatic FFD system and reliably keep it near unity to track that feature through a given video sequence.

3. RESTRICTED SCOPE FOR AVSP

As discussed in Section 2, accurate FFD is crucial to any AVSP system as it gives an upper bound on performance due to the *front-end effect*. FFD is a challenging task because of the inherent variability [2] as follows.

Pose: the images of a face vary due to the relative camera-face pose, with some facial features such as an eye or nose becoming partially or wholly occluded.

Presence or absence of structural components: facial features such as beards, mustaches, and glasses may or may not be present adding a great deal of variability in the appearance of a face.

Facial expression: a subject's face can vary a great deal due to the subject's expression (e.g., happy, sad, disgusted, and so on).

Occlusion: faces may be partially occluded by other objects.

Image orientation: facial features directly vary for different rotations about the camera's optical axis.

Imaging conditions: when captured, the quality of the image, and the facial features which exist within the image may vary due to lighting (spectra, source distribution and intensity) and camera characteristics (sensor response, lenses).

With over 150 reported approaches [2] to the field of face detection, the field is now becoming well established. Unfortunately, from all this research there is still no *one* technique that works best in all circumstances. Fortunately, the scope of the FFD task can be greatly narrowed due to the work in this paper, being primarily geared towards AVSP. For any AVSP application, the main visual facial feature of importance is the *mouth*. The extracted representation of the mouth does, however, require some type of normalisation for scale and rotation. It has been well documented [3] that the eyes are an ideal measure of scale and rotation of a face. To this end, FFD for AVSP will be restricted to eye and mouth detection.

To further simplify the FFD problem for AVSP, we can make the following number of assumptions about the images being processed:

(1) there is a single subject in each audio-visual sequence,
(2) the subject's facial profile is limited to frontal, with limited head rotation (i.e., ±10 degrees),
(3) subjects are recorded under reasonable (both intensity and spectral) lighting conditions,
(4) the scale of subjects remains relatively constant for a given video sequence.

These constraints are thought to be reasonable for most conceivable AVSP applications and are complied with in the M2VTS database [4] used throughout this paper for experimentation. Under these assumptions, the task of FFD becomes considerably easier. However, even under these less trying conditions, the task of accurate eye and mouth detection and tracking, so as to provide suitable normalisation and visual features for use in an AVSP application, is extremely challenging.

3.1. Validation

To validate the performance of an FFD system, a measure of relative error [3] is used, based on the distances between the expected and the estimated eye positions. The distance between the eyes (d_{eye}) has been long regarded as an accurate measure of the scale of a face [3]. Additionally, the detection of the eyes is an indication that the face search area does indeed contain a frontal face suitable for processing with an AVSP system. The distances d_l and d_r, for the left and right eyes, respectively, are used to describe the maximum distances between the true eye centers $\mathbf{c}_l, \mathbf{c}_r \in \mathbb{R}^2$ and the estimated positions $\hat{\mathbf{c}}_l, \hat{\mathbf{c}}_r \in \mathbb{R}^2$ as depicted in Figure 3.

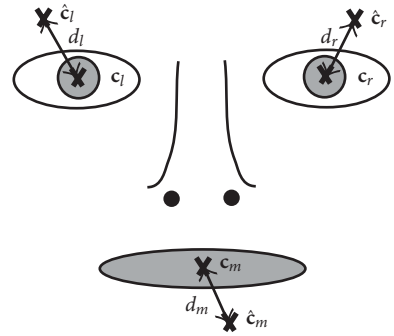


Figure 3: Relations between expected eye ($\mathbf{c}_l, \mathbf{c}_r$) and mouth ($\mathbf{c}_m$) positions and their estimated ones.

These distances are then normalised by dividing them by the distance between the expected eye centers ($d_{\text{eye}} = \|\mathbf{c}_l - \mathbf{c}_r\|$), making the measures independent of the scale of the face in the image and the image size,

$$e_{\text{eye}} = \frac{\max(d_l, d_r)}{d_{\text{eye}}} \tag{2}$$

The metric in (2) is referred to as the *relative eye error* e_{eye}. A similar measure is used to validate the performance of mouth location. A distance d_m is used to describe the distance between the true mouth position $\mathbf{c}_m \in \mathbb{R}^2$ and the estimated position $\hat{\mathbf{c}}_m \in \mathbb{R}^2$. This distance is then normalised by the distance between the expected eye centers, to make the measure also independent of the scale of the face in the image and the image size,

$$e_{\text{mouth}} = \frac{d_m}{d_{\text{eye}}}. \tag{3}$$

The metric in (3) is referred to as the *relative mouth error* e_{mouth}. Based on previous work by Jesorsky et al. [3], the eyes were deemed to be found if the relative eye error $e_{\text{eye}} < 0.25$. This bound allows a maximum deviation of half-an-eye width between the expected and estimated eye positions. Similarly, the mouth was deemed to be found if the relative mouth error $e_{\text{mouth}} < 0.25$.

All experiments in this paper were carried out on the audio-visual M2VTS [4] database, which has been used previously [5, 6] for AVSP work. The database used for our experiments consisted of 37 subjects (male and female) speaking four repetitions (shots) of ten French digits from *zero* to *nine*. For each speaker, the first three shots in the database, for the frames 1 to 100, had the eyes as well as the outer and inner labial contours, manually fitted at 10 frame intervals so as to gain the true eye and mouth positions. This resulted in over 1000 pretracked frames with 11 pretracked frames per subject per shot. The eye positions ($\mathbf{c}_l, \mathbf{c}_r$) were deemed to be

at the center of the pupil. The mouth position $\mathbf{c}_m$ was deemed to be the point of bisection on the line between the outer left and right mouth corners.

4. GAUSSIAN MIXTURE MODELS

A well-known classifier design which allows for modelling complex distributions parametrically are Gaussian mixture models (GMMs) [7]. Parametric classifiers have benefits over other classifiers as they give conditional density function estimates that can be directly applied to a Bayesian framework.

A GMM models the probability distribution of a statistical variable $\mathbf{x}$ as the sum of Q multivariate Gaussian functions

$$p(\mathbf{x}) = \sum_{i=1}^{Q} c_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\,|_{\mathbf{x}}, \tag{4}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})|_{\mathbf{x}}$ denotes a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and c denotes the mixture weight of class i. The parameters of the model $\boldsymbol{\lambda} = (c, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be estimated using the expectation maximization (EM) algorithm [8]. K-means clustering [9] was used to provide initial estimates of these parameters.

5. DEFINING THE FACE SEARCH AREA

The problem of FFD is a difficult problem due to the almost infinite number of manifestations nonfacial feature objects can take on in an input image. The problem of FFD can be greatly simplified if we are able to define an approximate face search area within the image. By searching within this face search area, the problem of eye and mouth detection can be greatly simplified due to the background being restricted to the face. This area of research is commonly referred to as *face segmentation*. Face segmentation can be defined as the segmenting of face pixels, usually in the form of a binary map, from the remaining background pixels in the image. Face segmentation approaches are excellent for defining a face search area as they aim at finding structural features of the face that exist even when the pose, scale, position, and lighting conditions of the face vary [2].

To gain this type of invariance, most face segmentation techniques use simplistic pixel or localised texture-based schemes to segment face pixels from their background. Techniques using simple grayscale texture measures have been investigated by researchers. Augusteijn and Skujca [10] were able to gain effective segmentation results by computing second-order statistical features on 16×16 grayscale subimages. Using a neural network, they were able to train the classifier using face and nonface textures, with good results reported. Human skin colour has been used and has proven to be one of the most effective pixel representations for face and skin segmentation [2]. Although different people have different skin colours, several studies have shown the major difference lies in the intensity, not chrominance representation, of the pixels [2, 11]. Several colour spaces have been explored for segmenting skin pixels [2] with most approaches adopting spaces in which the intensity component can be normalised or removed [11, 12]. Yang and Waibel [11] have achieved excellent segmentation results using normalised chromatic space $[r, g]$ defined in RGB (red, green, blue) space as

$$r = \frac{\mathrm{R}}{\mathrm{R}+\mathrm{G}+\mathrm{B}}, \qquad g = \frac{\mathrm{R}}{\mathrm{R}+\mathrm{G}+\mathrm{B}}. \tag{5}$$

It has been demonstrated in [11, 12] that once the intensity component of an image has been normalised, human skin obeys an approximately Gaussian distribution under similar lighting conditions (i.e., intensity and spectra). Under slightly differing lighting conditions, it has been shown that a generalised chromatic skin model can be generated using a mixture of Gaussians in a GMM. Fortunately, in most AVSP applications, it is possible to gain access to normalised chromatic pixel values from the face and background in training. It is foreseeable that, in most practical AVSP systems that have a stationary background, it would be possible to calibrate the system to its chromatic background through the construction of a chromatic background model when no subjects are present. Constructing an additional background GMM, segmentation performance can be greatly improved over the typical single hypothesis approach.

The task of pixel-based face segmentation using chromatic information can be formulated into the decision rule

$$\log p(\mathbf{o}_{rg}|\boldsymbol{\lambda}_{\text{skin}}) - \log p(\mathbf{o}_{rg}|\boldsymbol{\lambda}_{\text{back}}) \underset{\text{background}}{\overset{\text{skin}}{\lessgtr}} \text{Th}, \tag{6}$$

where Th is the threshold chosen to separate classes, with $p(\mathbf{o}_{rg}|\boldsymbol{\lambda}_{\text{skin}})$ and $p(\mathbf{o}_{rg}|\boldsymbol{\lambda}_{\text{back}})$ being used as the parametric GMM likelihood functions for the skin and background pixel classes in normalised chromatic space $\mathbf{o}_{rg} = [r, g]$. The prelabelled M2VTS database was employed to train up GMM models of the skin and background chromatic pixel values. Using the prelabelled eye coordinates and the distance between both eyes (d_{eye}), two areas were defined for training. The face area was defined as all pixels *within* the bounding box whose left and right sides are $0.5d_{\text{eye}}$ to the left of left eye x-coordinate and $0.5d_{\text{eye}}$ to the right of the right eye x-coordinate, respectively, with the top and bottom sides being $0.5d_{\text{eye}}$ above the average eye y-coordinate and $1.5d_{\text{eye}}$ below the average y-coordinate, respectively. The background area was defined as all pixels *outside* the bounding box whose left and right sides are d_{eye} to the left of the left eye x-coordinate and d_{eye} to the right of the right eye x-coordinate, respectively, with the top and bottom sides being d_{eye} above the average eye y-coordinate and the bottom of the input image, respectively. A graphical example of these two bounding boxes can be seen in Figure 4.

All prelabelled images from shot 1 of the M2VTS database were used in training $p(\mathbf{o}_{rg}|\boldsymbol{\lambda}_{\text{skin}})$ and $p(\mathbf{o}_{rg}|\boldsymbol{\lambda}_{\text{back}})$ GMMs. The GMMs were then evaluated on shots 2 and 3 of the M2VTS database achieving excellent segmentation in almost all cases. The skin GMM took on a topology of 8 diagonal mixtures with the background GMM taken on a

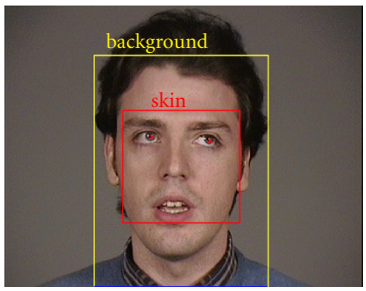


FIGURE 4: Example of bounding boxes used to gather skin and background training observations.

topology of 32 diagonal mixtures. The binary maps received after segmentation were then morphologically cleaned and closed to remove any spurious or noisy pixels. An example of the segmentation results can be seen in Figure 5.

6. AN APPEARANCE-BASED PARADIGM

In facial detection, there are a number of paradigms available. Techniques based on pixel, or texture-based segmentation, are useful for object location but do not provide any confidence on whether the object is there or not, making them less attractive for use in an object-detection capacity. Complicated iterative techniques such as active-shape models [13] or active-appearance models [14], which jointly model the intensity image variation and geometric form of the object, do provide such confidence measures but are quite computationally expensive. Appearance-based detection ignores the geometric form of the object completely and tries to model all variations in the object in terms of intensity value fluctuations within an ROI (window). In AVSP, this approach to FFD has an added benefit as recent research by Potamianos et al. [15] indicates that using simple intensity image-based representations of the mouth as input features perform better in the task of speechreading than geometric or joint representations of the mouth; indicating similar representations of the mouth may be used for detection and processing.

Appearance-based detection schemes work by sliding a 2D window $W(x, y)$ across an input image, with the contents of that window being classified as belonging to the object ω_{obj} or background ω_{bck} classes. The sliding of an $n_1 \times n_2$ 2D window $W(x, y)$ across an $N_1 \times N_2$ input image $I(x, y)$ can be represented as a concatenated matrix of vectors $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$, where the $D = n_1 n_2$ dimensional random vector $\mathbf{y}_t$ contains the vectorised contents of $W(x, y)$ centered at pixel coordinates (x, y). A depiction of this representation can be seen in Figure 6.

In reality, the concatenated matrix representation of $\mathbf{I}(x, y)$ is highly inefficient in terms of storage and efficiency of search, with the task of sliding a window across an image being far more effectively done through 2D convolution operations or a 2D FFT [16, 17]. However, the representation is used throughout this paper for explanatory purposes.

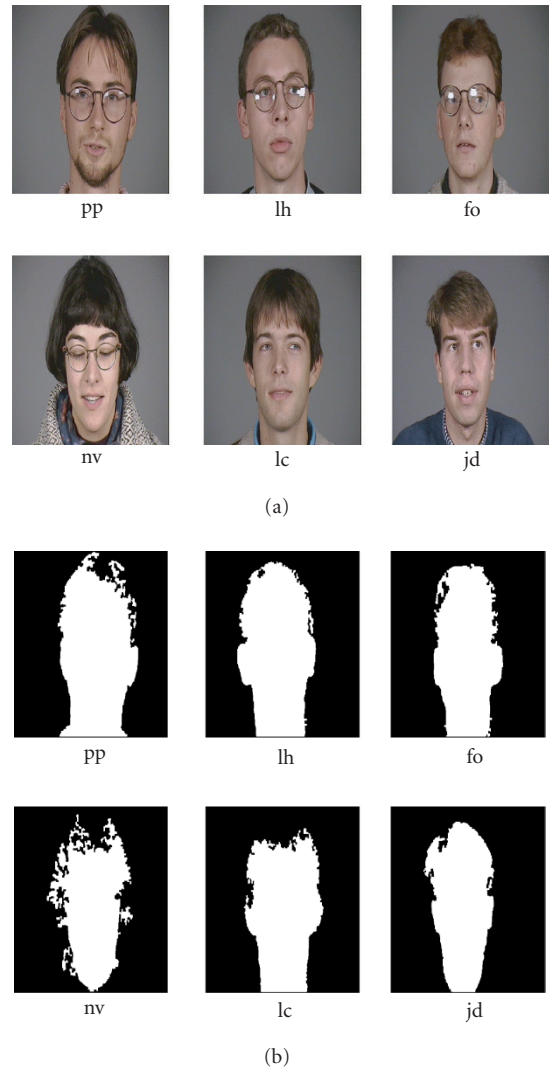


FIGURE 5: (a) Original example faces taken from the M2VTS database. (b) Binary potential maps generated using chromatic skin and background models.

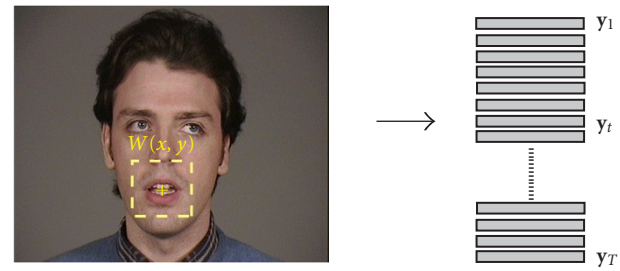


FIGURE 6: Demonstration of how contents of window $W(x, y)$ can be represented as vector $\mathbf{y}_t$.

The task of appearance-based object detection can be understood in a probabilistic framework as an approach

to characterise an object and its background as a class-conditional likelihood function $p(\mathbf{y}|\omega_{\text{obj}})$ and $p(\mathbf{y}|\omega_{\text{bck}})$. Unfortunately, a straightforward implementation of Bayesian classification is infeasible due to the high dimensionality of $\mathbf{y}$ and a lack of training images. Additionally, the parametric form of the object and background classes are generally not well understood. Hence, much of the work in an appearance-based detection concerns empirically validated parametric and nonparametric approximations to $p(\mathbf{y}|\omega_{\text{obj}})$ and $p(\mathbf{y}|\omega_{\text{bck}})$ [2].

6.1. Appearance-based detection framework

Any appearance-based detection scheme has to address two major problems:

(1) gaining a compact representation of $\mathbf{y}$ that maintains class distinction between object and background subimages, but is of small enough dimensionality to create a well-trained and computationally viable classifier;
(2) selection of a classifier to realise accurate and generalised decision boundaries between the object and background classes.

Most appearance-based object detection schemes borrow heavily on principal-component analysis (PCA) [18], or some variant, to generate a compact representation of the subimage $\mathbf{y}$. PCA is an extremely useful technique for mapping a D-dimensional subimage $\mathbf{y}$ into an M-dimensional subspace optimally in terms of reconstruction error. A fundamental problem with PCA is that it seeks a subspace that best represents a subimage in a sum-squared error sense. Unfortunately, in detection, the criteria for defining an M-dimensional subspace should be class separation between the object and background classes *not* reconstruction error. Techniques such as linear discriminant analysis (LDA) [18] produce a subspace based on such a criterion for detection [2, 18, 19, 20]. However, most of these techniques still require PCA to be used initially to provide a subspace that is free of any low-energy noise, that may hinder the performance of techniques like LDA [20, 21]. For this reason, most successful appearance-based detection schemes [2, 17] still use PCA or variant to some extent [22, 23, 24] to represent the subimage $\mathbf{y}$ succinctly.

The choice of what classifier to use in FFD is predominantly problem specific. The use of discriminant classifiers such as artificial neural networks (ANNs) [2] and support-vector machines (SVMs) [2, 25] has become prolific in recent times. ANNs and SVMs are very useful for classification tasks where the number of classes are static as they try to find the decision boundary directly for distinguishing between classes. This approach often has superior performance over parametric classifiers, such as GMMs, as parametric classifiers form their decision boundaries indirectly from their conditional class likelihood estimates. However, parametric classifiers, such as GMMs, lend themselves to more rigorous mathematical development and allow for the compact representation and classifier problems, associated with appearance-based detection, to be handled within the one framework. In this paper, GMMs are used to gain parametric likelihood functions $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}})$ and $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{bck}})$ for FFD experiments.

6.2. Single-class detection

PCA, although attractive as a technique for gaining a tractable likelihood estimate of $p(\mathbf{y})$ in a low-dimensional space, it does suffer from a critical flaw [22]. It does *not* define a proper probability model in the space of inputs. This is because the density is not normalised within the principal subspace. For example, if we were to perform PCA on some observations and then ask how well some *new* observations fit the model, the only criterion used is the squared distance of the new data from their projections into the principal subspace. An observation far away from the training observations, but nonetheless near the principal subspace, will be assigned a high "pseudo-likelihood" or low error. For detection purposes, this can have dire consequences if we need to detect an object using a single hypothesis test [18]. This is a common problem where the object class is well defined but the background class is not. This scenario can best be expressed as

$$l_1(\mathbf{y}) \underset{\omega_{\text{obj}}}{\overset{\omega_{\text{bck}}}{\lessgtr}} \text{Th}, \qquad l_1(\mathbf{y}) = \log\left[p(\mathbf{y}\,|\,\boldsymbol{\lambda}_{\text{obj}})\right], \tag{7}$$

where $l_1(\mathbf{y})$ is a score that discriminates between the object and background class with Th being the threshold for the decision. In this scenario, an object, which is drastically different in the true observation space, may be considered similar in the principal subspace or, as it will be referred to in this section, the *object space* (OS). This problem can be somewhat resolved by developing a likelihood function that describes both OS and its complementary *residual space* (RS). RS is referred to as the complementary subspace that is *not* spanned by the OS. Usually, this subspace cannot be computed directly, but a simplistic measure of its influence can be computed indirectly in terms of the reconstruction error realised from mapping $\mathbf{y}$ into OS. RS representations have proven exceptionally useful in single-hypothesis face detection. The success of RS representations in a single hypothesis can be realised in terms of energy. PCA naturally preserves the major modes of variance for an object in OS. Due to the background class not being defined, any residual variance can be assumed to stem from nonobject variations. Using this logic, objects with low-reconstruction errors can be thought more likely to stem from an object class rather than background class. Initial work by Turk and Pentland [16] used *just* the RS, as opposed to OS representation for face detection, as it gave superior results.

A number of approaches have been devised to gain a model to incorporate object and RS representations [16, 17, 19, 22, 23, 26] into $p(\mathbf{y}|\boldsymbol{\lambda})$. Moghaddam and Pentland [17] provided a framework for generating an improved representation of $p(\mathbf{y}|\boldsymbol{\lambda})$. In their work, they expressed the likelihood function $p(\mathbf{y}|\boldsymbol{\lambda})$ in terms of two independent Gaussian densities describing the object and residual spaces,

respectively,

$$p\left(\mathbf{y}|\boldsymbol{\lambda}^{\{\mathrm{OS+RS}\}}\right) = p\left(\mathbf{y}|\boldsymbol{\lambda}^{\{\mathrm{OS}\}}\right)p\left(\mathbf{y}|\boldsymbol{\lambda}^{\{\mathrm{RS}\}}\right), \tag{8}$$

where

$$p\left(\mathbf{y}|\boldsymbol{\lambda}^{\{\mathrm{OS}\}}\right) = \mathcal{N}\left(\mathbf{0}_{(M\times 1)}, \boldsymbol{\Lambda}_{(M\times M)}\right)\Big|_{\mathbf{x}}, \quad \mathbf{x} = \boldsymbol{\Phi}'\mathbf{y}, \tag{9}$$

$$p\left(\mathbf{y}|\boldsymbol{\lambda}^{\{\mathrm{RS}\}}\right) = \mathcal{N}\left(\mathbf{0}_{([\mathrm{R}-M]\times 1)}, \sigma^2\mathbf{I}_{([\mathrm{R}-M]\times[\mathrm{R}-M])}\right)\Big|_{\overline{\mathbf{x}}}, \quad \overline{\mathbf{x}} = \overline{\boldsymbol{\Phi}}'\mathbf{y} \tag{10}$$

such that $\boldsymbol{\Phi} = \{\phi_i\}_{i=1}^{M}$ are the eigenvectors spanning the subspace corresponding to the M largest eigenvalues λ_i, with $\overline{\boldsymbol{\Phi}} = \{\phi_i\}_{i=M+1}^{\mathrm{R}}$ being the eigenvectors spanning the residual subspace. The evaluation of (9) is rudimentary as it simply requires a mapping of $\mathbf{y}$ into the object subspace $\boldsymbol{\Phi}$. However, the evaluation of (10) is a little more difficult as we usually do not have access to the residual subspace $\overline{\boldsymbol{\Phi}}$ to calculate $\overline{\mathbf{x}}$. Fortunately, we can take advantage of the complementary nature of OS and the full observation space such that

$$\mathrm{tr}(\mathbf{Y}'\mathbf{Y}) = \mathrm{tr}(\boldsymbol{\Lambda}) + \sigma^2\,\mathrm{tr}(\mathbf{I}) \tag{11}$$

so that

$$\sigma^2 = \frac{\left[\,\mathrm{tr}(\mathbf{Y}'\mathbf{Y}) - \mathrm{tr}(\boldsymbol{\Lambda})\right]}{\mathrm{R}-M}, \tag{12}$$

allowing us to rewrite (10) as

$$p\left(\mathbf{y}|\boldsymbol{\lambda}^{\{\mathrm{RS}\}}\right) = \frac{\exp\left(-\epsilon^2(\mathbf{y})/2\sigma^2\right)}{\left(2\pi\sigma^2\right)^{(\mathrm{R}-M)/2}}, \quad \epsilon(\mathbf{y}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\boldsymbol{\Phi}\boldsymbol{\Phi}'\mathbf{y}, \tag{13}$$

where $\epsilon(\mathbf{y})$ can be considered as the error in reconstructing $\mathbf{y}$ from $\mathbf{x}$. This equivalence is possible due to the assumption of $p(\mathbf{y}|\boldsymbol{\lambda}^{\{\mathrm{RS}\}})$ being described by a Gaussian homoscedastic distribution (i.e., covariance matrix is described by an isotropic covariance $\sigma^2\mathbf{I}$). This simplistic isotropic representation of RS is effective, as the lack of training observations makes any other type of representation error prone. In a similar fashion to Cootes et al. [27], the ad hoc estimation of $\sigma^2 = (1/2)\lambda_{N+1}$ was found to perform best.

Many previous papers [17, 23, 24] have shown that objects with complex variations such as the mouth or eyes do not obey a unimodal distribution in their principal subspace. To model OS more effectively, a GMM conditional class likelihood estimate $p(\mathbf{y}|\boldsymbol{\lambda}^{\{\mathrm{OS}\}})$ was used to account for these complex variations. The same ensemble subimages that were used to create the eigenvectors spanning OS were used to create the GMM density estimate. An example of this complex clustering can be seen in Figure 7 where multiple mixtures have been fitted to the OS representation of an ensemble of mouth subimages.

Similar approaches have been proposed for introducing this residual in a variety of ways such as factor analysis (FA) [19], sensible principal-component analysis (SPCA) [22], or probabilistic principal-component analysis (PPCA) [23]. For the purposes of comparing different detection metrics,

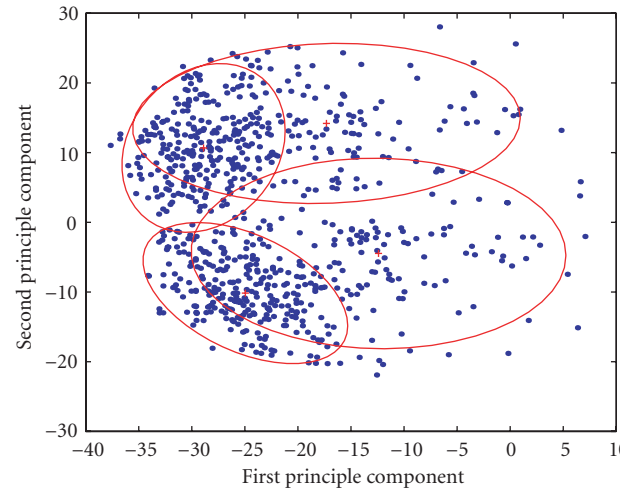


Figure 7: Example of multimodal clustering of mouth subimages within principal subspace.

the experimental work presented in this paper concerning the combining of OS and RS subimage representations will be constrained to the complementary approach used by Moghaddam and Pentland [17].

6.3. Two-class detection

As discussed in the previous section, the use of RS, or more specifically reconstruction error, can be extremely useful when trying to detect an object when the background class is undefined. A superior approach to detection is to have well defined likelihood functions for the object and background classes. The two-class detection approach can be posed as

$$l_2(\mathbf{y}) \underset{\omega_{\mathrm{obj}}}{\overset{\omega_{\mathrm{bck}}}{\lessgtr}} \mathrm{Th},$$
$$l_2(\mathbf{y}) = \log\left[p(\mathbf{y}|\boldsymbol{\lambda}_{\mathrm{obj}})\right] - \log\left[p(\mathbf{y}|\boldsymbol{\lambda}_{\mathrm{bck}})\right]. \tag{14}$$

A problem presents itself in how to gain observations from the background class to train $\boldsymbol{\lambda}_{\mathrm{bck}}$. Fortunately, for FFD, the face area is assumed to be approximately known (i.e., from the skin map), making the construction of a background model plausible as the type of nonobject subimages is limited to those on the face and surrounding areas. Estimates of the likelihood functions $p(\mathbf{y}|\boldsymbol{\lambda}_{\mathrm{obj}})$ and $p(\mathbf{y}|\boldsymbol{\lambda}_{\mathrm{bck}})$ can be calculated using GMMs, but we require a subspace that can adequately discriminate between the object and background classes. To approximate the object and background likelihood functions, we could use the original OS representation of $\mathbf{y}$. Using OS for building parametric models, we may run the risk of throwing away vital discriminatory information, as OS was constructed under the criterion of optimally reconstructing the object not the background. A more sensible approach is to construct a *common space* (CS) that adequately reconstructs both object and background subimages.

A very simple approach is to create a CS using roughly the same number of training subimages from both the object and

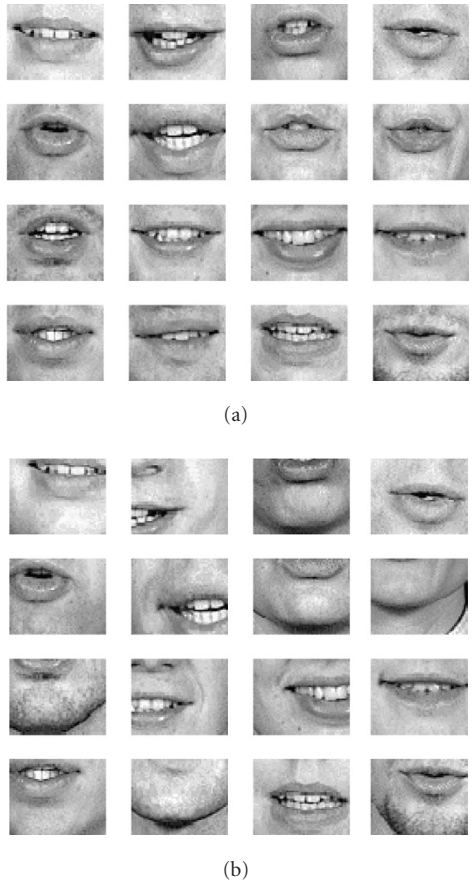


Figure 8: Example of (a) mouth subimages, (b) mouth background subimages.

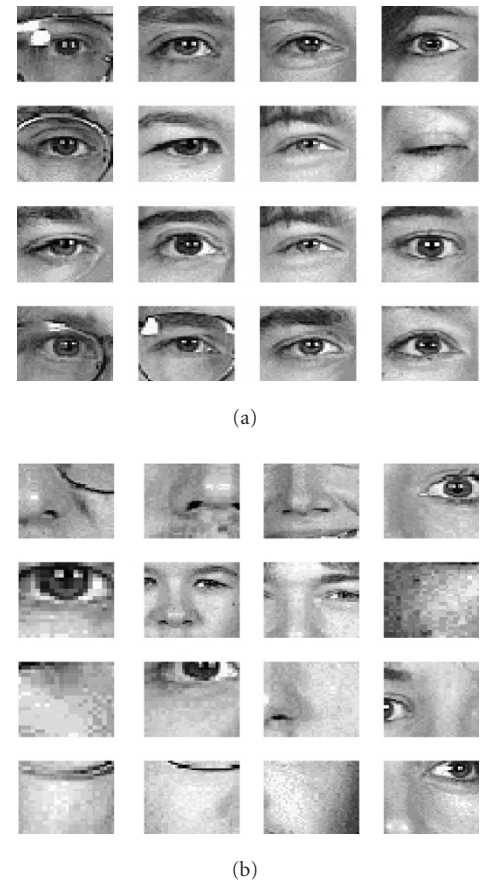


Figure 9: Example of (a) eye subimages, (b) eye background subimages.

background classes. A problem occurs in this approach as there are far more background subimages than object subimages per training image. To remedy this situation, background subimages were selected randomly during training from around the object in question. An example of randomly selected mouth, mouth background, eye, and eye background subimages can be seen in Figures 8 and 9, respectively. Note for the eye background subimages in Figure 9b that the scale varies as well. This was done to make the eye detector robust to a multiscale search of the image.

As previously mentioned, PCA is suboptimal from a discriminatory standpoint as the criterion for gaining a subspace is reconstruction error not class separability. LDA can be used to construct a *discriminant space* (DS) based on such a criterion. Since there are only two classes ($L = 2$) being discriminated between (i.e., object and background), LDA dictates that DS have a dimensionality of one, due to the rank being restricted to $L - 1$. This approach would work well if both the object and background classes were described adequately by a single Gaussian, each with the same covariance matrix. In reality, we know that this is rarely the case with eye, mouth, and background distributions being modelled far more accurately using multimodal distributions. Using this knowledge, an intraclass clustering approach can be employed to build a DS by describing both the object and background distributions with several unimodal distributions of approximately the same covariance.

The technique can be described by defining $\mathbf{Y}_{\text{obj}}$ and $\mathbf{Y}_{\text{bck}}$ as the training subimages for the object and background classes. Principal subspaces $\mathbf{\Phi}_{\text{obj}}$ of size M_{obj} and $\mathbf{\Phi}_{\text{bck}}$ of size M_{bck} are first found using normal PCA. The object subspace $\mathbf{\Phi}_{\text{obj}}$ and background subspace $\mathbf{\Phi}_{\text{bck}}$ are found separately to ensure that most discriminative information is preserved while ensuring any low-energy noise that may corrupt LDA in defining a suitable DS is removed. A joint orthonormal base $\mathbf{\Phi}_{\text{jnt}}$ is then found by combining object and background subspaces via the Gram-Schmidt process. The final size of $\mathbf{\Phi}_{\text{jnt}}$ is constrained by M_{obj} and M_{bck} and the overlap that exists between object and background principal subspaces. The final size of the joint space is important as it needs to be as low as possible for successful intraclass clustering whilst preserving discriminative information. For experiments conducted in this paper, successful results were attained by setting M_{obj} and M_{bck} to 30.

Soft clustering was employed to describe each class with several approximately equal covariance matrices. K-means

clustering [9] was first employed to gain initial estimates of the clusters with the EM algorithm, then refining the estimates. For the experiments conducted in this paper, best performance was attained when 8 clusters were created from the compactly represented object subimages $\mathbf{Y}_{\text{obj}}\mathbf{\Phi}_{\text{jnt}}$ and 16 clusters created from the compactly represented background subimages $\mathbf{Y}_{\text{obj}}\mathbf{\Phi}_{\text{jnt}}$. This resulted in a virtual $L = 24$ class problem resulting in a 23 $(L-1)$-dimensional DS after LDA. Once DS was found, estimates of $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{DS}\}})$ and $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{bck}}^{\{\text{DS}\}})$ were calculated normally using a GMM.

6.4. Evaluation of appearance models

In order to have an estimate of detection performance between object and nonobject subimages $\mathbf{y}$, the prelabelled M2VTS database was employed to evaluate performance for eye and mouth detection. In training and testing, illumination invariance was obtained by normalising the subimage $\mathbf{y}$ to a zero-mean unit-norm vector [17].

A very useful way to evaluate detection performance of different appearance models is through the use of detection-error trade-off (DET) curves [28]. DET curves are used as opposed to traditional receiver-operating characteristic (ROC) due to their superior ability to easily observe performance contrasts. DET curves are used for the detection task, as they provide a mechanism to analyse the trade-off between missed detection and false alarm errors.

Results are presented here for the following detection metrics.

OS-L1: object space representation of $\mathbf{y}$ for the single hypothesis score $l_1(\mathbf{y})$ where $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{OS}\}})$ is approximated by an 8-mixture diagonal GMM. OS is a 30-dimensional space.

OS-L2: object space representation of $\mathbf{y}$ for the two-class hypothesis score $l_2(\mathbf{y})$ where $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{OS}\}})$ is an 8-mixture diagonal GMM and $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{bck}}^{\{\text{OS}\}})$ is a 16-mixture diagonal GMM. OS is a 30-dimensional space.

RS-L1: residual space representation of $\mathbf{y}$ for the single hypothesis score $l_1(\mathbf{y})$ where $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{RS}\}})$ is parametrically by single mixture isotropic Gaussian. The OS used to gain the RS metric was a 5-dimensional space.

OS+RS-L1: complementary object and RS representation of $\mathbf{y}$ for the single hypothesis score $l_1(\mathbf{y})$ where $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{OS+RS}\}}) = p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{OS}\}})p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{RS}\}})$. The likelihood function $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{OS}\}})$ is parametrically described by a 8-mixture diagonal GMM, with $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{RS}\}})$ being described by single mixture isotropic Gaussian. OS is a 5-dimensional space.

CS-L2: common space representation of $\mathbf{y}$ for the two-class hypothesis score $l_2(\mathbf{y})$ where $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{CS}\}})$ is an 8-mixture diagonal GMM and $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{bck}}^{\{\text{CS}\}})$ is a 16-mixture diagonal GMM. CS is a 30-dimensional space.

DS-L2: discriminant space representation of $\mathbf{y}$ for the two-class hypothesis score $l_2(\mathbf{y})$ where $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{obj}}^{\{\text{DS}\}})$ is an 8-mixture diagonal GMM and $p(\mathbf{y}|\boldsymbol{\lambda}_{\text{bck}}^{\{\text{DS}\}})$ is a 16-mixture diagonal GMM. DS is a 23-dimensional space.

The same GMM topologies were found to be effective for both mouth and eye detection. In all cases, classifiers were trained using images from shot 1 of the M2VTS database with testing being performed on shots 2 and 3. To generate DET curves for eye and mouth detection, 30 random background subimages were extracted for every object subimage. In testing, this resulted in over 5000 subimages being used to generate DET curves, indicating the class separation between object and background classes. As previously mentioned, the eye background subimages included those taken from varying scales to gauge performance in a multiscale search. Both the left and right eyes were modeled using a single model. Figure 10 contain DET curves for the eye and mouth detection tasks, respectively.

Inspecting Figure 10, we can see the OS-L1 metric performed worst overall. This can be attributed to the lack of a well-defined background class and the OS representation of subimage $\mathbf{y}$ not giving sufficient discrimination between object and background subimages. Performance improvements can be seen from using the reconstruction error for the RS-L1 metric, with further improvement being seen in the complementary representation of subimage $\mathbf{y}$ in the OS+RS-L1 metric. Note that a much smaller OS was used (i.e., $M = 5$) for the OS+RS-L1 and RS-L1 metrics to ensure that the majority of object energy is contained in OS and the majority of background energy is in RS. It can be seen that *all* the single hypothesis L1 metrics have poorer performance than any of the L2 metrics, signifying the large performance improvement gained from defining an object and background likelihood function. There is some benefit in using the CS-L2 metric over the OS-L2 metric for both eye and mouth detection. The use of the DS-L2 metric gives the best performance over all metrics in terms of equal error rate.

Figure 10 are only empirical measures of separability between the object and background classes for various detection metrics. The true measure of object detection performance can be found in the actual act of detecting an object in a given input image. For the task of eye detection, each top-left half and top-right half of the skin map is scanned with a rectangular window to determine whether there is a left and right eye present. A depiction of how the skin map is divided for FFD can be seen in Figure 11.

A location error metric first presented by Jesorky et al. [3] and elaborated upon in Section 3.1 for eye detection was used in our experiments; this metric states that the eyes are deemed to be detected if both the estimated left and right eye locations are within 0.25 deye of the true eye positions. To detect the eyes at different scales, the input image and its skin map was repeatedly subsampled by a factor of 1.1 and scanned for 10 iterations with the original scale chosen so that the face could take up 55% of the image width. Again, tests were carried out on shots 2 and 3 of the prelabelled M2VTS database. The eyes were successfully located at a rate of 98.2% using the DS-L2 metric. A threshold was employed from DET analysis to allow for a false alarm probability of 1.5%, which in term resulted in only 13 false alarms over the 700 faces tested. The use of this threshold was very important as it gave an indication of whether the eyes, and subsequently an accurate measure of scale, had been found for locating the mouth.

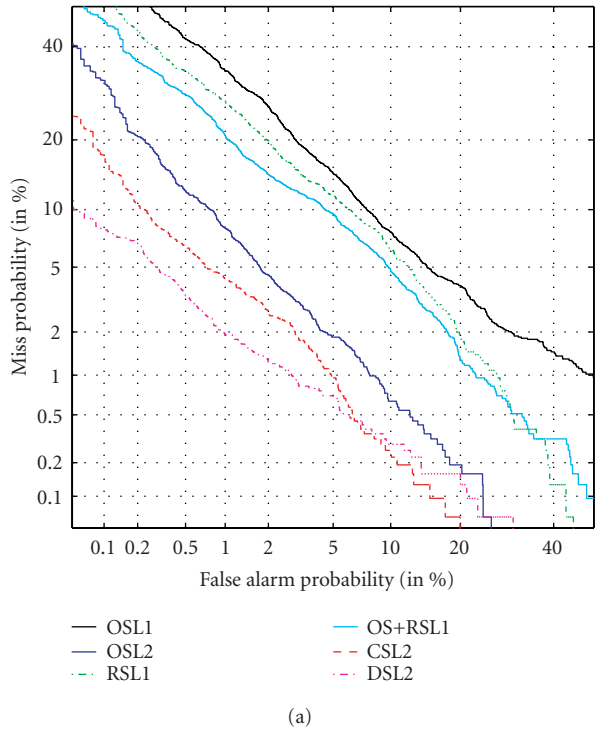


Figure 10: DET curve of different detection metrics for separation of (a) eyes and (b) mouth between background subimages.

Given that the scale of the face is known (i.e., distance between the eyes d_{eye}), the mouth location performance was tested on shots 2 and 3 of the prelabelled M2VTS database.

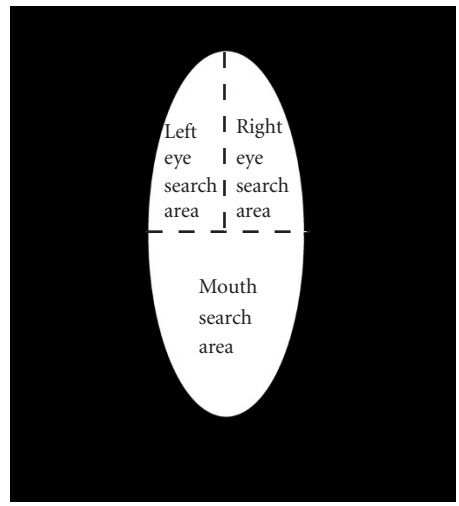


Figure 11: Depiction of how skin map is divided to search for facial features.

The lower half of the skin map is scanned for the mouth, with a mouth being deemed to be located if the estimated mouth center is within $0.25d_{\text{eye}}$ of the true mouth position. The mouth was successfully detected at a rate of 92.3% using the DS-L2 metric. When applied to the task of tracking in a continuous video sequence, this location rate starts approaching 100% due to the smoothing of the mouth coordinates through time via a median filter.

7. DISCUSSION

Appearance-based detection of the eyes and mouth is of real benefit in AVSP applications. The appearance-based paradigm allows for detection, not just location, which is essential for effective AVSP applications. A number of techniques have been evaluated for the task of appearance-based eye and mouth detection. All techniques differ primarily in their representation of the subimage $\mathbf{y}$ being evaluated and how an appropriate likelihood score is generated. Techniques based on single-class detection (similarity measure based solely on the object) have been shown to be inferior to those generated from two-class detection (similarity measure based on both the object and background classes). Similarly, the need for gaining a compact representation of the subimage $\mathbf{y}$ that is discriminatory between the mouth and background is beneficial, as opposed to approaches that generate a compact representation of the object or both classes based on reconstruction error.

A technique for creating a compact discriminant space has been outlined using knowledge of LDA's criterion for class separation. In this approach, an intraclass clustering approach is employed to handle the typical case of when both the object and background class distributions are multimodal. Using this approach, good results, suitable for use in AVSP, were achieved in practice for the tasks of eye detection and mouth detection.

REFERENCES

[1] F. Lavagetto, "Converting speech into lip movements: a multimedia telephone for hard-of-hearing people," *IEEE Trans. on Rehabilitation Engineering*, vol. 3, no. 1, pp. 90–102, 1995.

[2] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[3] O. Jesorsky, K. Kirchberg, and R. Frischholz, "Robust face detection using the Hausdorff distance," in *3rd Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, pp. 90–95, Halmstad, Sweden, June 2001.

[4] S. Pigeon, "The M2VTS database," Tech. Rep., Laboratoire de Télécommunications et Télédétection, Université Catholique de Louvain, Louvain, Belgium, 1996.

[5] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[6] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[7] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 639–643, 1994.

[8] A. P. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[9] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic, Dordrecht, The Netherlands, 1992.

[10] M. F. Augusteijn and T. L. Skujca, "Identification of human faces through texture-based feature recognition and neural network technology," in *Proc. IEEE International Conference on Neural Networks*, vol. I, pp. 392–398, Piscataway, NJ, USA, 1993.

[11] J. Yang and A. Waibel, "A real-time face tracker," in *Proc. 3rd IEEE Workshop on Applications of Computer Vision*, pp. 142–147, Sarasota, Fla, USA, 1996.

[12] M.-H. Yang and N. Ahuja, "Detecting human faces in color images," in *Proc. IEEE International Conference on Image Processing*, vol. 1, pp. 127–130, Chicago, Ill, USA, 1998.

[13] J. Luettin, N. A. Thacker, and S. W. Beet, "Speechreading using shape and intensity information," in *Proc. International Conf. on Spoken Language Processing*, vol. 1, pp. 58–61, Philadelphia, Pa, USA, October 1996.

[14] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. A. Bangham, "Lipreading using shape, shading and scale," in *Auditory-Visual Speech Processing*, vol. 1, pp. 73–78, Sydney, Australia, 1998.

[15] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. IEEE International Conference on Image Processing*, vol. 3, pp. 173–177, Chicago, Ill, USA, 1998.

[16] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[17] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.

[18] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, London, UK, 2nd edition, 1990.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley and Sons, New York, NY, USA, 2nd edition, 2001.

[20] M.-H. Yang, N. Ahuja, and D. Kriegman, "Face detection using mixtures of linear subspaces," in *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 70–76, Grenoble, France, March 2000.

[21] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[22] S. Roweis, "EM algorithms for PCA and SPCA," in *Neural Information Processing Systems*, vol. 10, pp. 626–632, Denver, Col, USA, 1997.

[23] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," Tech. Rep. NCRG/97/010, Neural Computing Research Group, Aston University, Birmingham, UK, September 1997.

[24] B. Chalmond and S. C. Girard, "Nonlinear modeling of scattered multivariate data and its application to shape change," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 422–432, 1999.

[25] Y. Li, S. Gong, J. Sherrah, and H. Liddell, "Multi-view face detection using support vector machines and eigenspace modelling," in *Proc. 4th International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, pp. 241–244, Brighton, UK, August 2000.

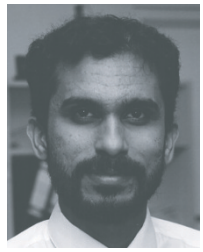
[26] D. J. Bartholomew, *Latent Variable Models and Factor Analysis*, Charles Griffin & Co., London, UK, 1987.

[27] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "The use of active shape models for locating structures in medical images," *Image and Vision Computing*, vol. 12, no. 6, pp. 355–365, 1994.

[28] A. Martin, G. Doddington, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech Conference*, vol. 4, pp. 1895–1898, Rhodes, Greece, September 1997.

Simon Lucey received the B.Eng. (Hons) degree from the University of Southern Queensland, Toowoomba, Australia, in 1998. In January 1999, he joined the Research Concentration in Speech, Audio, and Video Technology at the Queensland University of Technology (QUT), Brisbane, Australia, where he is completing his Ph.D. His research interests are in the fields of audio-visual speech/speaker recognition, classifier combination, visual feature extraction for visual speech processing and FFD. Mr. Lucey is a student member of the Institute of Electrical and Electronic Engineers.

Sridha Sridharan obtained his B.S. degree in electrical engineering and M.S. degree in communication engineering from the University of Manchester's Institute of Science and Technology, UK, and the Ph.D. degree in signal processing from the University of New South Wales, Australia. Dr. Sridharan is a senior member of the IEEE, USA, and a corporate member of the IEE, the United Kindgom, and The Institution of Engineers, Australia (IEAust). He is currently a Professor in the School of Electrical and Electronic Systems Engineering of the Queensland University of Technology (QUT) and is also the Head of the Research Concentration in Speech, Audio, and Video Technology at QUT.



Vinod Chandran received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 1982, the M.S. degree in electrical engineering from Texas Tech University, Lubbock, Texas, USA, in 1985, and the Ph.D. degree in electrical and computer engineering and the M.S. degree in computer science from Washington State University, Pullman, Washington, USA, in 1990 and 1991, respectively. He is currently a Senior Lecturer at the Queensland University of Technology, Brisbane, Australia, in the School of Electrical and Electronic Systems Engineering. He is a member of the Research Concentration in Speech, Audio, and Video Technology. His research interests include pattern recognition, high-order spectral analysis, speech processing, image processing, audio and video coding, and content-based retrieval. Dr. Chandran is a member of the IEEE and ACM, and a member of Tau Beta Pi and Phi Kappa Phi.